УДК 519.235

РАЗРАБОТКА ВИЗУАЛЬНОГО МЕТОДА ИССЛЕДОВАНИЯ ЗАВИСИМОСТИ КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ НА ОСНОВЕ ТАБЛИЦ СОПРЯЖЕННОСТИ

***Ольга Александровна Бакаева**, старший преподаватель*
Тел. 8 917 697 8233, e-mail: helga_rm@rambler.ru
ФГБОУ ВПО Мордовский государственный университет им. Н.П. Огарева
http://www.math.mrsu.ru

В данной статье описан визуальный метод исследования зависимости категориальных переменных, основанный на геометрической интерпретации частот в таблице сопряжённости. Произведена дифференциация независимости на идеальную, статистическую и практическую.

Ключевые слова: категориальные переменные, таблицы сопряжённости, частоты, зависимость, визуальный метод, идеальная независимость, статистическая независимость, практическая независимость.

Введение

Особенности традиционного подхода к исследованию проблемы зависимости переменных требует чётко определённого вероятностного пространства и случайного эксперимента. В действительности же в большинстве случаев встречается эксперимент, характеризующийся конечным числом условно упорядоченных значений переменной. Поэтому для любого исследователя важны не числовое выражение степени зависимости и его обоснование, а более грубые понятия – «практически независимы», «практически зависимы», а также вид зависимости – «возрастающая», «убывающая». В связи с этим существует необходимость разработки численных методов исследования зависимости, способных практически оценивать наличие связи.


О.А. Бакаева

В результате использования различных методов исследования зависимости происходит анализ и выбор факторов, т.е. отличительных особенностей объектов. При обработке такого рода информации особую роль играют категориальные переменные, то есть те, которые принимают качественные значения, и бинарные переменные с двумя альтернативными значениями. Такие переменные встречаются достаточно часто в технических, социальных и биомедицинских системах, что обуславливает необходимость их исследования. Эффективным, наглядным и универсальным инструментом обработки таких данных являются таблицы сопряжённости [1].

Для категориальных переменных существует достаточно много способов выявления связи с помощью аппарата таблиц сопряжённости, но среди них нет универсального. В связи с этим в зависимости от расположения и значений частот приходится использовать тот или иной критерий проверки связи.

Методы, основу которых составляет вычислительный этап с последующим анализом полученной статистики, относят к аналитическим. Это первые критерии оценки связи, коэффициент отношения избытка, G-критерий Вульфа, точный критерий Фишера, коэффициенты взаимной сопряжённости К. Пирсона и А. Чупрова, шанс и шансовое отношение, различные модификации классического критерия χ^2 проверки независимости. К методам, при реализации которых необходимо использование ЭВМ и специальных программ, относят: G-критерий Вульфа (MS Excel), шанс (MS Excel, Калькулятор таблиц сопряжённости), критерий Фишера (Fisher Exact), пакет анализа статистических данных «STATISTICA» [2]. Как видно, критериев исследования связи между категори-

альными переменными достаточно много. Но, к сожалению, многие из них неприменимы в силу малости частот или нулевого значения частоты в одной из ячеек. Поэтому в данной области существует проблема исследования связи для таких «особенных» таблиц сопряжённости. Необходим наглядный подход, который бы позволил не только с помощью расчётов, но и визуально определять наличие зависимости и таким образом повысил уровень достоверности и обоснованности выводов о её наличии.

Теоретические моменты

Метод визуализации зависимости основан на графической интерпретации данных из таблицы сопряжённости. Пусть имеются две переменные A и B, предположим, что обе они категориальные, к тому же бинарные, т.е. принимают по два различных значения A_1, A_2 и B_1, B_2 соответственно. Тогда таблица частот 2 × 2 будет иметь вид:

Таблица 1.

Схема таблицы сопряженности 2 × 2

	B_1	B_2	Всего
A_1	f_{11}	f_{12}	f_{10}
A_2	f_{21}	f_{22}	f_{20}
Всего	f_{01}	f_{02}	f_{00}

Исходя из геометрического смысла строк и столбцов и последующего построения прямых, можно делать выводы о независимости бинарных категориальных переменных.

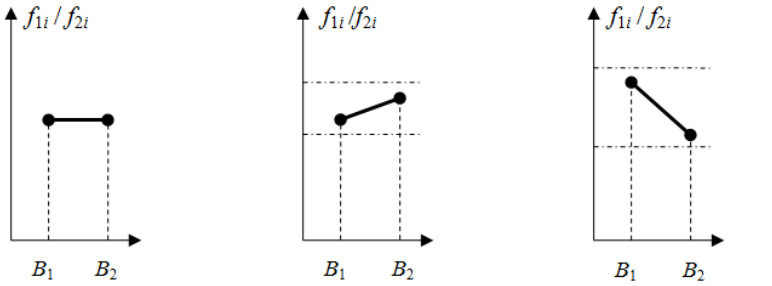
Отличительная особенность данного метода состоит в том, что все известные критерии являются расчётно-аналитическими, т.е. в своей основе они содержат вычисления различных статистик, далее следует сравнение вычисленных статистик с критическими значениями и последующие выводы. Однако достаточно часто аналитическое решение можно представить графически. В случае исследования зависимости между бинарными категориальными переменными, визуальный метод позволяет не только выявить сам факт отсутствия связи между переменными, но и получить оценку уровня значимости или незначимости.

В прикладном программном продукте «STATISTICA» можно увидеть графическое представление любой таблицы сопряжённости через график взаимодействия частот. Предлагается реализовать аналогичный подход к выявлению связи между бинарными переменными в MS Excel.

Сам визуальный метод исследования независимости состоит в следующем: для каждой таблицы сопряжённости строится одна прямая, и по ее расположению относительно горизонтальной оси OB, делается вывод о независимости исследуемых категориальных переменных. Данная прямая будет проходить через две точки, которые имеют координаты

$$\left(B_1, \frac{f_{11}}{f_{21}}\right) \text{ и } \left(B_2, \frac{f_{12}}{f_{22}}\right), \text{ при условии, что и } f_{21} \neq 0 \text{ и } f_{22} \neq 0.$$

Если переменные независимы, то получается горизонтальная прямая, в случае зависимости прямая – наклонная, знак наклона соответствует знаку зависимости: при положительной зависимости прямая убывает, при отрицательной – возрастает. При наличии нулевых ячеек можно использовать поправку Йетса или поменять изображаемую переменную. Однако практически прямая никогда не будет строго горизонтальной из-за целочисленности экспериментальных частот и наличия случайного разброса наблюдений. Поэтому необходимо использование специальных критериев для установления значимости отличия прямой от горизонтальной – рис. 1.



Идеальная независимость (не реализуется)

Статистическая независимость (статистический критерий)

Практическая независимость (незначимость)

Рис. 1. Графическое представление различных видов независимости

Можно указать диапазон расположения прямых, при котором будет статистическая независимость, то есть в рамках случайного разброса невозможно отличить наблюдаемую таблицу от идеально независимой. Это соответствует выполнению статистического критерия, например, χ^2. Можно также указать диапазон расположения, при котором будет лишь слабая зависимость, которой с практической точки зрения можно пренебречь (практическая независимость). Применение этого критерия обосновывается:

а) указанной выше нечёткостью определения случайного эксперимента, т.е. отсутствие гарантий однородности полученных частот;

б) в практике использования категориальных переменных, которые сами недостаточно четко определены, слабая зависимость действительно незначима и обычно отбрасывается.

Величины, которые откладываются на оси ординат, обозначим за критерии $z_i = \frac{f_{1i}}{f_{2i}}$. Тогда условие практической незначимости может быть введено через коэффициент корреляции:

$$\left| \frac{f_{11} f_{22} - f_{12} f_{21}}{\sqrt{f_{01} f_{02} f_{10} f_{20}}} \right| < \rho_0 ,$$

где ρ_0 – пороговое значение коэффициента корреляции, ниже которого зависимость считается незначимой. Отсюда получается условие на величину разности критериев

$$|z_1 - z_2| < \rho_0 \left| \frac{\sqrt{f_{01} f_{02} f_{10} f_{20}}}{f_{22} f_{21}} \right| .$$

Для конкретных значений частот эту разность можно показать на графике как предельные границы, что и будет визуально определять незначимость зависимости.

Для проверки статистической незначимости можно воспользоваться критерием χ^2 без предварительного расчета теоретической частоты:

$$\frac{f_{00} \ (f_{11} f_{22} - f_{12} f_{21})^2}{f_{10} f_{20} f_{01} f_{02}} < {\chi_T}^2 .$$

Если сделать замену $\rho_0 = \sqrt{\frac{{\chi_T}^2}{f_{00}}}$, то это условие легко визуализируется, и можно говорить о наличии зависимости с определённым уровнем незначимости.

Из всего вышесказанного следует вывод, что предложенный метод визуализации зависимости между бинарными категориальными переменными имеет интерпретацию с различных точек зрения. Его преимущества заключаются в том, что он позволяет делать выводы о различных уровнях независимости, а с помощью порогового коэффициента корреляции можно визуально определять незначимость зависимости.

Практическая часть

Задача 1. В Республике Мордовия в 2009-2010гг. было зарегистрировано 75 случаев заболевания вирусом А (H1N1) («свиной грипп») среди взрослых людей (18-58 лет) и 45 случаев среди детей (0-17 лет). Среди взрослых было 9 летальных исходов, среди детей летальных исходов не было. Частоты результатов лечения выглядят следующим образом.

Таблица 2

Таблица частот результата лечения вируса А (H1N1) среди взрослых и детей в Республике Мордовия в 2009-2010гг.

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0-17 лет	0	45	45
Взрослые 18-58 лет	9	66	75
Всего	9	111	120

Задача состоит в исследовании связи между категориальными переменными «возраст» и «результат лечения», т.е. нужно ответить на вопрос, влияет ли возраст на результат лечения, и если влияет, то в какой степени – как сильно.

Из таблицы следует, что среди детей летальных исходов не наблюдалось, т.е. все 45 детей выздоровели (100%). Среди взрослых летальный исход наблюдался в 9 случаев из 75, а 66 человек выздоровели (88%). Нужно доказать статистически, что выздоровление носит неслучайный характер и понять, значимо ли отличается количество выздоровевших детей и взрослых.

Задача 2. В Республике Мордовия в 2010-2011гг. было зарегистрировано 254 пациента, заболевших вирусом А (H1N1) («свиной грипп»). Из них в г.о. Саранск проживает 195 человек, в районах республики – 59 человек. В г.о. Саранск заболевание в 4-х случаях закончилось летальным исходом. По районам зарегистрировано 2 смертельных исхода.

Таблица 3

Таблица частот результата лечения вируса А (H1N1)09 в зависимости от места проживания зимой 2010-2011гг.

	Результат лечения		
Место проживания	Летальный	Выздоровление	Всего
Проживающие в г.о. Саранск	4	191	195
Проживающие в др. городах и районах	2	57	59
Всего	6	248	254

Проверим наличие связи между переменными «место проживания» и «результат лечения», т.е. зависит ли результат лечения от места проживания пациента.

Расчеты

Задача 1. Исходя из данных табл.2, переменная *A* – это «возраст», переменная *B* – «результат лечения». Чтобы выявить зависимость между этими переменными необхо-

димо построить прямую, проходящую через точки ($x_1 = B_1$ = «Дети»; $y_1 = \frac{f_{11}}{f_{21}}$) и A_2 ($x_2 = B_2$ = «Взрослые»; $y_2 = \frac{f_{12}}{f_{22}}$). Изобразим расположение этой прямой на графике (рис. 2).

Полученную прямую можно считать практически горизонтальной (строго горизонтальной она будет только в случае «идеальной» независимости, такое на практике не встречается), поэтому можно сделать вывод о статистической независимости.

Рис.2. Прямая, характеризующая взаимодействие переменных «Возраст» – «Результат лечения»

Для проверки статистической независимости воспользуемся критерием χ^2 без предварительного расчета теоретической частоты:

$$\frac{120\cdot(0\cdot 66-9\cdot 45)^2}{45\cdot 75\cdot 9\cdot 111}=5{,}84\,.$$

Так как полученное значение 5,84 < 5,99 = χ^2_T (0,05; 2), то можно сделать вывод о статистической незначимости зависимости.

Для оценки практической независимости рассчитаем величину

$$\left|\frac{f_{11}f_{22}-f_{12}f_{21}}{\sqrt{f_{01}f_{02}f_{10}f_{20}}}\right|=\left|\frac{0\cdot 66-9\cdot 45}{\sqrt{45\cdot 75\cdot 9\cdot 111}}\right|=0{,}22\,.$$

Примем коэффициент корреляции ρ_0 = 0,05. Так как полученное значение 0,22 > 0,05, то зависимость между факторами не является незначимой.

Из аналитических методов самым точным считается критерий Фишера [3]. Значение данного коэффициента P = 0,012, что можно трактовать следующим образом: гипотеза о независимости переменных отвергается с достоверностью 0,95 (на уровне знач. α = 0,05). Как видно, аналитический метод подтверждает выводы, полученные с помощью визуального критерия.

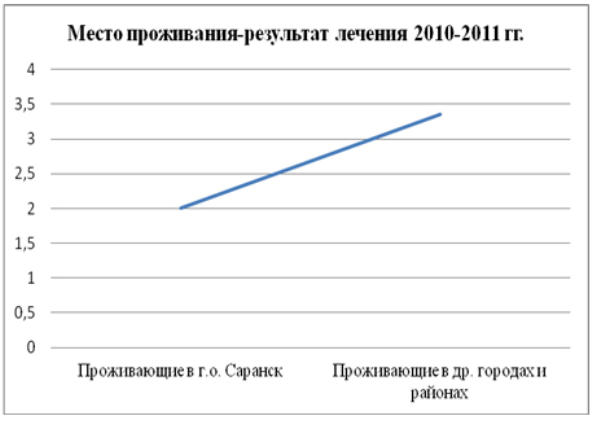


Рис.3. Прямая характеризующая взаимодействие переменных «Возраст» – «Результат лечения»

Таким образом, переменные «возраст» и «результат лечения» зависимы (статистически), т.е. для различных возрастных групп населения эффективность лечения тоже различна.

Задача 2. Исходя из данных табл.3, переменная A – это «место проживания», переменная B – «результат лечения». Для исследования зависимости между этими переменными необходимо построить прямую, проходящую через точки ($x_1 = B_1$ = «Проживающие в г.о. Саранск»; $y_1 = \frac{f_{11}}{f_{21}}$) и A_2 ($x_2 = B_2$ = «Проживающие в др. городах и районах»; $y_2 = \frac{f_{12}}{f_{22}}$). Прямая будет выглядеть следующим образом:

Полученная прямая визуально имеет отклонение от горизонтального положения. Проверим взаимодействие переменных, используя критерий χ^2 без предварительного расчёта теоретической частоты:

$$\frac{254\cdot(4\cdot 57-2\cdot 191)^2}{195\cdot 59\cdot 6\cdot 248}=0{,}35\,.$$

Так как полученное значение 0,35 < 5,99 = χ^2_T (0,05; 2), то можно сделать вывод о статистической незначимости зависимости.

Для оценки практической независимости переменных рассчитаем величину

$$\left|\frac{f_{11}f_{22}-f_{12}f_{21}}{\sqrt{f_{01}f_{02}f_{10}f_{20}}}\right|=\left|\frac{4\cdot 57-191\cdot 2}{\sqrt{195\cdot 59\cdot 6\cdot 248}}\right|=0{,}04\,.$$

Примем коэффициент корреляции ρ_0 = 0,05. Так как полученное значение 0,04 < 0,05, то зависимость считается незначимой.

Сравним полученный вывод с результатом, вычисленным только аналитически с помощью точного критерия Фишера. Статистика данного критерия P = 0,424, что говорит о независимости факторов с вероятностью 0,95 (на уровне знач. α = 0,05). Как видно, результаты аналогичны.

Для переменных «место проживания (лечения)» и «результат лечения» наличие связи не определено, т.е. эти факторы независимы (практически). Отсюда вывод: эффективность лечения в городском округе Саранск и районах республики одинакова и не зависит от места проживания (лечения) пациента.

Заключение

Автор считает, что в данной работе новыми являются следующие положения и результаты:

– визуальный метод исследования зависимости категориальных переменных, позволяющий, исходя из расположения прямой отношения категорий относительно горизонтальной оси, делать выводы о различных уровнях независимости;

– предложены геометрическая интерпретация различных уровней независимости – идеальной, статистической и практической;

– разработаны аналитические критерии и статистики, позволяющие в совокупности с геометрической интерпретацией независимости, получать достоверные практические выводы о взаимодействии переменных;

– произведено сравнение выводов о наличии связи, полученных визуальным методом и классическим критерием Фишера, которое свидетельствует о работоспособности и достоверности предложенного визуального метода;

– решены практические задачи по выявлению связи между категориальными переменными «возраст», «место проживания» и «результат лечения» заболевания вирусом A (H1N1) – «свиной грипп».

Литература

1. *Бакаева О.А.* Оценка связи между качественными признаками с помощью таблиц сопряженности / Р. Р. Бикмурзина, О.А. Бакаева, А.А. Панина // Технические и естественные науки: проблемы, теория, практика: межвуз. сб. науч. тр. Вып. X. – Саранск: РНИИЦ, 2009. С. 33-38.

2. *Бакаева О.А.* Алгоритм выбора рационального способа проверки наличия зависимости между категориальными переменными при донозологическом контроле // Информационные технологии моделирования и управления. 2013. № 1 (79). С. 4-11.

3. *Бакаева О.А.* Использование точного критерия Фишера для выявления связи между категориальными переменными // XL Огаревские чтения: материалы науч. конф. – Саранск: Изд-во Мордов. ун-та, 2012. С. 154-57.

Development a visual method of research of dependence of the categorical variables based on contingency tables

Ol'ga Alexanrovna Bakaeva, senior teacher
The Mordovian State University named N. P. Ogarev

In this article a visual method of research of dependence of the categorical variables, based on geometrical interpretation of frequencies in the contingency table is described. Differentiation of independence is produced on ideal, statistical and practical.

Keywords: categorical variables, contingency tables, frequency, dependence, visual method, ideal independence, statistical independence, practical independence.

УДК 51-76

МОДЕЛИРОВАНИЕ ФОРМЫ СЕМЯН ЛИМСКОЙ ФАСОЛИ (Phaseolus limensis L.)

***Ирина Семеновна Виноградова**, старший научный сотрудник,*
профессор кафедры физики
Тел.: 3912 494 678, e-mail: vis.akadem@mail.ru
Сибирский государственный технологический университет, Красноярск, Россия
www.kit-sibstu.ru

Форма семян является важной их характеристикой, необходимой при проектировании оборудования для сбора урожая на плантациях, транспортировки, сушки, хранения и проветривания высушенных семян. В работе проведены расчёты объемов и площадей поверхности семян лимской фасоли в процессе их созревания из измеренных размеров - длины, ширины и толщины. Для расчётов применялись модели, используемые в литературе, проводится сравнение с экспериментально измеренными значениями объёма.

Ключевые слова: семена, лимская фасоль, выращивание, измерения и расчёты объёма и площади поверхности.

В последние годы физические свойства семян интенсивно исследовались в зарубежной литературе, их результаты опубликованы в ведущих журналах по технологии продуктов питания. В этих работах изучались размеры семян, их масса, объем и площадь поверхности, пористость, отклонение от сферичности и другие физические свойства. Эта информация важна не только для инженеров, проектирующих машины, но и для растениеводов. Использование компьютерной техники позволяет расширить диапазон измерений и проводить аттестацию качества и классификацию сорта продукта. Считается, что внешний вид продукта является важным для маркетинга и продажи. Размер, форма, цвет и наличие пятен и дефектов влияет на восприятие покупателя.


И.С. Виноградова

Фасоль Лима происходит из Южной Америки. Свое название она получила от города Лима в Перу, где ее обнаружили европейцы. Ее научное название PhaseoluslimensisL. или PhaseoluslunatusL. Это второй по значению американский вид фасоли, которую из-за согнутых бобов и сплющенных семян называют лунообразной. Лимскую фасоль в основном возделывают в районах жаркого климата - в Центральной и Южной Америке, на Антильских островах, в Африке, тропической Азии. Она культивируется в южных районах России - на Северном Кавказе, в Закавказье, в Молдове, перспективная культура для Средней Азии и юга Украины. Её семена содержат около 20% белка, который представлен в основном глобулинами и альбуминами, около 60% крахмала, 1,6-1,9% жира. Полкило Лимы содержит столько же питательных веществ, сколько содержится в 1 кг мяса.

В настоящей работе проведены измерения на семенах лимской фасоли основных геометрических параметров: длины, ширины и толщины, проведены расчёты их объё-